Journal of Machine Learning Research X (200X) X-X Submitted X/XX; Published XX/XX

An Information Criterion for Variable Selection in Support Vector Machines

Gerda Claeskens GERDA.CLAESKENS@ECON.KULEUVEN.BE
Christophe Croux CHRISTOPHE.CROUX@ECON.KULEUVEN.BE
Johan Van Kerckhoven JOHAN.VANKERCKHOVEN@ECON.KULEUVEN.BE
ORSTAT and Leuven Statistics Research Center
Katholieke Universiteit Leuven
B-3000 Leuven, Belgium

Editor: Isabelle Guyon and Amir Reza Saffari Azar Alamdri

Abstract

Support vector machines for classification have the advantage that the curse of dimensionality is circumvented. It has been shown that a reduction of the dimension of the input space leads to even better results. For this purpose, we propose two information criteria which can be computed directly from the definition of the support vector machine. We assess the predictive performance of the models selected by our new criteria and compare them to existing variable selection techniques in a simulation study. The simulation results show that the new criteria are competitive in terms of generalization error rate while being much easier to compute. We arrive at the same findings for comparison on some real-world benchmark datasets.

Keywords: Information Criterion, Supervised Classification, Support Vector Machine, Variable Selection.

1. Introduction

We study classification using the support vector machine (SVM). We start from a training set $\{(x_i, y_i)\}$ containing n observations. Each p-dimensional observation $x_i = (x_{i1}, \ldots, x_{ip})$ has a class label y_i assigned to it, which is either $+1$ or -1. We wish to find a function $f(\cdot)$ such that for an observation x the predicted class $\hat{y} = +1$ if $f(x)$ is positive, and $\hat{y} = -1$ if $f(x)$ is negative. We want this function to assign the correct class labels to the training observations (low training error rate) and to accurately classify new observations (low generalization error rate). Working with a subset of the p variables $x_{i1}, \ldots, x_{ip}$ reduces variability of the class-label estimator and might lead to better out-of-sample predictions.

It is only true to some extent that variable selection would not be necessary in the support vector machine setting since it manages to circumvent the so-called "curse of dimensionality" (see for example Cristianini and Shawe-Taylor, 2000, Hastie, Tibshirani, and Friedman, 2001, or Schölkopf and Smola, 2002). While the SVM approach avoids fitting a number of parameters equal to the dimension of the input space, there remains the high probability of a perfect separation in high-dimensional problems. For example, if p is larger than the number of observations, it is always possible to perfectly separate the two classes of training data by a hyperplane. In general, the risk of overfitting will increases with the

dimension for most data configurations. Hence, the risk of obtaining a decision rule with poor generalization properties (high generalization error rate) cannot be avoided. Guyon et al. (2002) illustrate this and show that variable selection can further improve the SVM's performance.

Variable selection techniques can be divided into three categories. Filters select subsets of variables as a pre-processing step, independently of the prediction method. Wrappers utilize the classification method to score subsets of variables. Finally, embedded methods include variable selection into the construction of the classifier. In this paper we propose new information criteria for SVMs, yielding a wrapper method where we consider the SVM merely as a black box. We refer to Guyon and Elisseeff (2003) for an introduction to variable and feature selection in Machine Learning. Information criteria are a standard tool for model selection in traditional statistics. Information criteria for variable selection assign a numerical value to each subset of the variables under consideration. The subset with the lowest value of the information criterion is then selected. Examples are the Akaike information criterion (AIC, Akaike, 1973) and the Bayesian information criterion (BIC, Schwarz, 1978). Claeskens and Hjort (2008) survey and explain the use of common information criteria for statistical variable selection in likelihood-based models, we refer to there for more references.

For support vector machines only very few information criteria have been developed. The kernel regularisation information criterion (KRIC) of Kobayashi and Komaki (2006) was originally proposed for parameter tuning of the SVM. We apply it for variable selection. However, the KRIC has a complicated definition and is computationally expensive for large sample sizes. In this paper two new information criteria are proposed, one shares properties with AIC, the other with BIC. We want the new criteria to select a preferably compact subset of variables with good predictive properties. We will show that submodels selected by the new criteria are as performing as the ones chosen by the KRIC, while they incur substantially less computational overhead. We also make a comparison with using cross-validated error rate based criteria, as in Kearns et al. (1997). An important contribution of this paper is that our numerical comparisons show that the popular, but time consuming, cross-validation criteria are outperformed in generalization error by the new information criteria, where the latter are coming at almost no additional computational cost.

Alternative approaches perform variable selection in feature space instead of in input space (Shih and Cheng, 2005), or select a set of "maximally separating directions" in the input space Fortuna and Capson (2004). These methods, however, do not select a set of original input variables. Various other authors have suggested different formulations for the SVM such that variable selection is performed automatically. Examples of such embedded methods can be found in Bi et al. (2003), Zhu et al. (2004), Neumann, Schnörr and Steidl (2005), Lee et al. (2006), Wang, Zhu, and Zou (2006), Zhang (2006), and Lin and Zhang (2006).

In Section 2 we define the support vector machine setting, we review existing information criteria and we describe ranking techniques to speed up the variable selection process. In Section 3, we define the new information criteria and highlight their advantages. Section 4 contains the results of a simulation study and in Section 5 we compare the different techniques on a few real-world benchmark datasets. Section 6 concludes and gives some directions for further research.

2. Problem Setting

2.1 The Support Vector Machine

We denote the training sample (x_i, y_i), $1 \leq i \leq n$, with x_i a p-dimensional vector of explicative variables, and $y_i \in \{-1, +1\}$ the class label. The goal is to estimate a target function $f(x)$ in the space of explicative variables such that $f(x_i) > 0$ for $y_i = +1$, and $f(x_i) < 0$ for $y_i = -1$.

We start with linear support vector machines, where $f(x)$ is of the form $f(x) = w'x + b$. For binary classification this function is obtained by solving the minimisation problem

$$\min_{w,b,\xi_i} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right\} \text{ subject to } \begin{cases} y_i(w'x_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, \ldots, n. \end{cases} \tag{1}$$

The ξ_i are slack margin variables, indicating how close a point x_i lies to the separating boundary (if $\xi_i < 1$), or how badly it is misclassified (if $\xi_i > 1$). The tuning parameter C controls how much weight is put on trying to achieve perfect separation.

The dual problem can be solved more easily, and has the following form:

$$\min_{\alpha}\{\frac{1}{2}\alpha' Q\alpha - \sum_{i=1}^{n} \alpha_i\} \text{ subject to } \begin{cases} 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n, \\ \sum_{i=1}^{n} y_i\alpha_i = 0. \end{cases} \tag{2}$$

Here α_i is the weight given to the observation (x_i, y_i), and Q is a positive semi-definite matrix with entries $Q_{i,j} = y_i y_j x_i' x_j$. The vector w can be found from $w = \sum_{i=1}^{n} y_i\alpha_i x_i$. The negative intercept b is found by computing $b = 0.5(r_2 - r_1)$, where

$$r_1 = \frac{\sum_{0<y_i\alpha_i<C}(Q\alpha)_i - 1}{\sum_{0<y_i\alpha_i<C} 1} \text{ and } r_2 = \frac{\sum_{0>y_i\alpha_i>-C}(Q\alpha)_i - 1}{\sum_{0>y_i\alpha_i>-C} 1}.$$

If no i exist for which $0 < y_i\alpha_i < C$, then define

$$r_1 = \frac{1}{2}\Big(\min_{\alpha_i=0, y_i=1}(Q\alpha)_i - \max_{\alpha_i=C, y_i=1}(Q\alpha)_i \Big),$$

and analogously for r_2, with $y_i = -1$. Note that we can write $\xi_i = [1 - y_i a_i]_+$, where $[x]_+ = \max\{0, x\}$ and where $a_i = f(x_i)$.

The linear SVM can be extended towards more complex decision functions in a rather straightforward way. Therefore we replace the inner products $x_i' x_j$ in the definition of Q by a more general kernel function $K(x_i, x_j)$. See Cristianini and Shawe-Taylor (2000) for the properties that these kernel functions must have. This leads to a more general decision function

$$f(x) = \sum_{i=1}^{n} y_i\alpha_i K(x_i, x) + b. \tag{3}$$

Popular choices for the kernel function in (3) are the linear kernel, where the kernel function is $K(x, z) = x'z$, the polynomial kernel of the form $K(x, z) = (c_0 + \gamma x'z)^d$, and the radial basis kernel $K(x, z) = \exp(-\gamma\|x - z\|^2)$, where c_0, γ and d are regularization parameters

that can be tuned for optimal performance of the classifier. In this more general setting, we have

$$\|w\|^2 = \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) = \alpha' Q \alpha$$

for the squared norm of the weight vector, where $Q_{i,j} = y_i y_j K(x_i, x_j)$.

2.2 Existing Variable Selection Techniques for SVM

We compare our new methods (Section 3) to variable selection based on (ten-fold) cross-validation (CV), guaranteed risk minimisation (GRM, Vapnik 1982) and the kernel regularisation information criterion (KRIC) by Kobayashi and Komaki (2006). Each of these will be explained in more detail below.

Ten-fold cross-validation divides the training data in ten parts of roughly equal size. One part is left out, the other nine parts are the training data and are used to fit the SVM. This SVM is applied to the part that is left out to obtain an estimate of the error rate. This process is repeated ten times (each time a different part is left out) to obtain the CV generalization error rate $\hat{\varepsilon}(S)$ as the average of the ten separate error rates. We select the model with the lowest value of $\hat{\varepsilon}(S)$, where S ranges over all subsets of variables under consideration. Another common method is five-fold CV. The lower the number of folds, the less computing time is required, but the higher the variability of the estimates of the generalization error. Note that $n-$fold CV is the same as the computationally infeasible leave-one-out CV.

General risk minimisation (Vapnik, 1982) is derived from the estimated generalization error rate, using

$$GRM(S) = \hat{\varepsilon}(S) + \frac{|S|}{n}\big(1 + \sqrt{1 + \hat{\varepsilon}(S)(n/|S|)}\big). \tag{4}$$

Here, $|S|$ stands for the number of input variables in the set S and n is the number of observations in the training sample. We select the model with the lowest value of $GRM(S)$, where S ranges over all subsets of variables under consideration. Kearns et al. (1997) compare CV, GRM and minimum description length (Rissanen, 1989). Their experiments have demonstrated that none of the criteria is consistently better than the others. Note that the computational overhead for computing these measures can be immense, since we need to train ten support vector machines to estimate the generalization error rate for only one submodel.

We now define the KRIC of Kobayashi and Komaki (2006). This criterion was originally developed to tune the constant C in the SVM definition (1), and by extension to tune the kernel parameters. We use it without much adjustment for variable selection. Denote by $x_{i,S}$ the subvector of x_i, consisting of elements x_{ij} with $j \in S$, and similarly for other vectors. We estimate the SVM (1) using the observations $(x_{i,S}, y_i)$, yielding the vectors ω_S, b_S and ξ_S, where the subscript S refers to the subset of variables under consideration. In the dual problem (2), we have $\alpha_S = (\alpha_{S,1}, \ldots, \alpha_{S,n})$ and $[Q_S]_{i,k} = y_i y_k K(x_{i,S}, x_{k,S})$. The decision rule $f_S(x)$ is as in (3), and we set $a_{i,S} = f_S(x_{i,S})$. Next, we define vectors t_S and m_S of length n, with components

$$t_{S,i} = \eta^2 \frac{\exp(-\eta a_{i,S} y_i)}{(1 + \exp(-\eta a_{i,S} y_i))^2} \quad \text{and} \quad m_{S,i} = -\eta \frac{y_i \exp(-\eta a_{i,S} y_i)}{1 + \exp(-\eta a_{i,S} y_i)}, \quad i = 1, \ldots, n.$$

Here we choose $\eta = \log(2)$ such that $\log(1+\exp(-\eta x))$ and $\eta[1-x]_+$ coincide for $x = 0$, see Kobayashi and Komaki (2006) for further motivation. With $\lambda = C^{-1}\log 2$ the KRIC for the logistic Bayesian model for SVMs is defined as

$$\begin{aligned} \mathrm{KRIC}(S) &= 2\Bigg[\sum_{i=1}^{n}\log\left(1+\exp(-\eta a_{i,S}y_i)\right) \\ &\quad + \mathrm{trace}((Q_S\mathrm{diag}(t_S)+\lambda I_n)^{-1}Q_S(\mathrm{diag}(m_S)^2 - n^{-1}m_S m_S^t))\Bigg]. \end{aligned} \tag{5}$$

Alternatively, Sollich's Bayesian model for SVMs (Sollich, 2002) leads to a KRIC with a similar form as the one in (5). Using

$$\nu(a_{i,S}) = (1+\exp(-2C))^{-1}(\exp(-C[1-a_{i,S}]_+) + \exp(-C[1+a_{i,S}]_+)),$$

the KRIC for the Sollich Bayesian model for SVMs is defined as

$$\mathrm{KRICS}(S) = \mathrm{KRIC}(S) - 2n\log\sum_{i=1}^{n}\nu(a_{i,S}). \tag{6}$$

The computation of the KRIC includes inverting an $n \times n$-matrix with only a few zeroes. Therefore, the computation is time-consuming if the sample size n is large. Both the CV error rate and the KRIC may require a prohibitive computing time when a large number of different models needs to be evaluated.

2.3 Ranking Techniques

A full subset search is computationally not feasible even not for problems with only a small number of dimensions ($p = 15$ for example). To dramatically reduce the number of models while still selecting a model that is "almost" the best model, Chen, Li and Li (2005) use a genetic algorithm, while Peng, Long and Ding (2005) suggest a combined backward elimination/forward selection strategy. However, both of these techniques still suffer from the possibility that a large number of models needs to be checked before arriving at a solution.

Alternatively, variable ranking consists of assigning a "value of importance" to each variable and sorting the variables according to their importance. This results in a series of p stacked models, thus only p evaluations of the variable selection criterion are needed. The most commonly used algorithm is the SVM recursive feature elimination (SVM-RFE) technique from Guyon et al. (2002). For a linear SVM, the variables are ranked by w_j^2, with w_j the j-th component of the weight vector w. This technique assumes that the variables are standardized to have mean 0 and variance 1. The extension proposed by Rakotomamonjy (2003) allows application to SVMs with a non-linear kernel. We use the following SVM-RFE algorithm with variable influence

$$\Delta\|w_S\|^2_{(j)} = \left|\|w_S\|^2 - \|w_{S\setminus\{j\}}\|^2\right|$$

as suggested by Rakotomamonjy (2003).

Step 1: Initialise $S \leftarrow \{1, \ldots, p\}$, the subset of unranked features, and $r \leftarrow ()$, the vector of ranked features.

Step 2: Repeat the following steps until $S = \emptyset$.

(a) Train a SVM on $(x_{i,S}, y_i)$, and compute $\|w_S\|^2 = \alpha_S' Q_S \alpha_S$.

(b) For each $j \in S$, train a new SVM on $(x_{i,S\setminus\{j\}}, y_i)$. This gives a value $\|w_{S\setminus\{j\}}\|^2 = \alpha'_{S\setminus\{j\}} Q_{S\setminus\{j\}} \alpha_{S\setminus\{j\}}$ for each $j \in S$.

(c) Obtain $j_0 = \text{argmin}_j \, |\|w_S\|^2 - \|w_{S\setminus(j)}\|^2|$ and set $S \leftarrow S \setminus \{j_0\}$ and $r \leftarrow (j_0, r)$.

The vector r contains the ranked variables, with the first element the most important one. A disadvantage of this method is that the number of SVMs to be trained is $\mathcal{O}(p^2)$. This can be overcome by using α_S instead of $\alpha_{S\setminus\{j\}}$ in Step 2*b*, such that $\|w_{S\setminus\{j\}}\|^2 \approx \alpha_S' Q_{S\setminus\{j\}} \alpha_S$. Rakotomamonjy (2003) argues that this will not affect the ranking significantly, while still allowing a major reduction in computational time, bringing the number of SVMs to be estimated to $\mathcal{O}(p)$. We employ this approximation in the simulation study in Section 4 and in the real data examples in Section 5.

The most easiest way to rank the variables is by filtering methods. Zhang et al. (2006) propose using $s_j = |w_j(m_{j,+1} - m_{j,-1})|$ for ranking, where $m_{j,+1}$ and $m_{j,-1}$ are the within-class means of variable j. Shih and Cheng (2005) use the Fisher score

$$S_j = \frac{|m_{j,+1} - m_{j,-1}|}{\sqrt{\sigma_{j,+1}^2 + \sigma_{j,-1}^2}}$$

for a linear SVM, where $\sigma_{j,+1}^2$ and $\sigma_{j,-1}^2$ are the within-class variances of variable j. The main advantage of using S_j is that it is not necessary to train any SVM to rank the variables. The Fisher score ranking is considered in Sections 4 and 5.

3. The New Information Criteria

As stated in the previous section, evaluating the CV error rate or the KRIC of a particular support vector machine model requires a high number of additional computations. For this reason, we propose two new criteria which use information already available in the SVM, without additional complicated computations. The criteria are based on how badly the SVM violates the margin constraints, which are written as $\sum_{i=1}^n \xi_{i,S}$, where $\xi_{i,S}$ is the margin slack of observation i in the support vector machine trained on the variables with indices in S, where S is a subset of $\{1, \ldots, p\}$. Alternatively, we can use the logarithm of this sum, analogous to Bai and Ng (2002) for selecting the number of factors in factor analysis. However, in the SVM setting this has the drawback that the value is undefined if the sum equals zero, which can happen if the data are perfectly separable. Also, Bai and Ng (2002) advise using a log-transform for scalar invariance reasons. Since we follow the advice to standardise the variables before training the SVM, for better ranking as explained in Section 2.3, we automatically have scalar invariance of the sum of the margin slacks. For these reasons, we choose not to take the log-transform.

Generally (but not always), $\sum_i \xi_{i,S}$ will decrease as more variables are added. Therefore we add a penalty term related to the number of included variables to ensure a tradeoff

between accuracy and simplicity of the chosen model. We suggest adding a linear penalty term, such that we get an information criterion of the form

$$IC(S) = \sum_{i=1}^{n} \xi_i + C(n)|S|\,, \tag{7}$$

where S is the set of variables included in the model.

A first choice is to take $C(n)$ constant in (7). It is interesting to note that $IC(S)$ is then, up to constant factors, an easily computable approximation of the KRIC of Kobayashi and Komaki (2006), hereby providing a theoretical justification for its use. To better understand this, note first that $\log\left(1+\exp(-\eta a_{i,S} y_i)\right)$ is a continuous approximation of the hinge loss function $\eta[1 - y_i a_{i,S}]_+ = \eta\xi_{i,S}$ for all $1 \leq i \leq n$. Hence, the first term of the KRIC can be approximated, up to a constant factor, by $\sum_i \xi_{i,S}$. For the approximation of the second term in (5), rewrite

$$\begin{aligned} W &= (Q_S \mathrm{diag}(t_S) + \lambda I_n)^{-1} Q_S (\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t) \\ &= V \mathrm{diag}(t_S)^{-1} (\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t), \end{aligned}$$

with $V = (A + \lambda I_n)^{-1} A$ a symmetric, positive semi-definite matrix and $A = Q_S \mathrm{diag}(t_S)$. Denoting A^- the generalised inverse of A, and using a series expansion around $\lambda = 0$, gives that the leading term of $V = A^-(I + \lambda A^-)^{-1} A$ is equal to $A^- A$. This expansion converges as long as the eigenvalues of λA^- are strictly less than one, which can be obtained by taking λ small enough. We now use a singular value decomposition of both A and A^- and use the fact that the singular values of A^- are the reciprocals of the non-zero singular values of A, to obtain that the product $A^- A$ is a $n \times n$ diagonal matrix with on the diagonal $|S|$ ones and the remaining entries zero. Thus, the leading term of trace(W) equals the sum of $|S|$ diagonal entries of the matrix $\mathrm{diag}(t_S)^{-1}(\mathrm{diag}(m_S)^2 - n^{-1} m_S m_S^t)$. The i-th diagonal element of this matrix is equal to

$$\frac{n-1}{n} t_{S,i}^{-1} m_{S,i}^2 = \frac{n-1}{n} \exp(-\eta a_{i,S} y_i).$$

To further facilitate computations we replace this by 1, motivated by the fact that $\eta a_{i,S} y_i$ is often small. Although this approximation might be crude for a single term, we found empirically that it works well for the summation over the entire training set. Hence, we arrive at the approximation trace$(W) \approx |S|$ which is the linear penalty term in (7).

Taking the constant value $C(n) = 2$, leads to our first new support vector machine information criterion (SVMIC):

$$\mathrm{SVMICa}(S) = \sum_{i=1}^{n} \xi_i + 2|S|. \tag{8}$$

The newly proposed criterion SVMICa for support vector machines shares the form of the penalty with the well-known Akaike (1973) information criterion. This AIC is defined as minus twice the value of the maximised log likelihood of the model, plus two times the number of parameters to be estimated (that is, $2|S|$). Because the penalty $2|S|$ is not dependent on the sample size n, we expect that both criteria share some properties, such as having

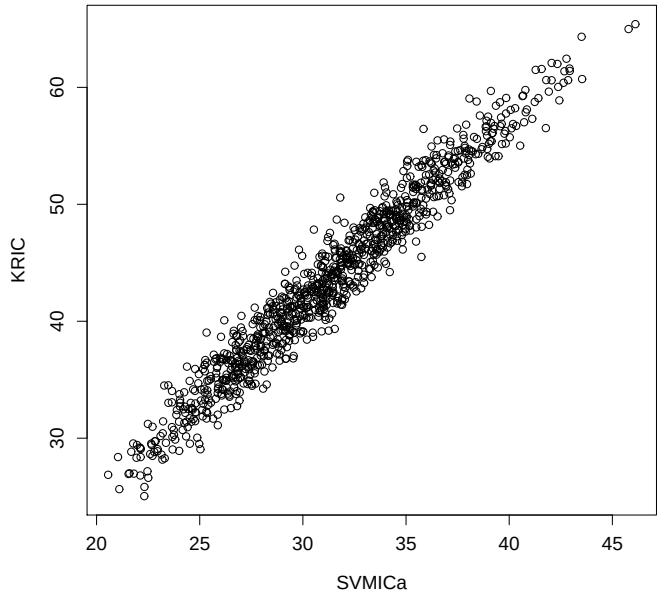

Figure 1: Values of KRIC and SVMICa in a simulation experiment, showing high correlation (0.975).

the tendency to not select the most parsimonious model. For the AIC, Woodroofe (1982) has shown that in the limit for $n \to \infty$, the expected number of superfluous parameters is less than one.

To support the definition of SVMICa , we ran a simulation experiment and compared the values of KRIC and SVMICa for 100 models. The sample size is $n = 50$, with 10 variables of which only the first 4 variables are different from zero. A detailed description of the simulation setting can be found in Section 4. We used a linear kernel. Figure 1 reports these numerical results and shows a high correlation (0.975) between the values of the two criteria. Other simulation settings gave comparable correlation values.

Our second proposed criterion follows the spirit of Schwarz's (1978) Bayesian information criterion (BIC). This criterion is defined similarly as the AIC, but instead of the penalty $2|S|$, it uses $\log(n)|S|$. The BIC has been shown to be consistent (Haughton 1988, 1989). This means that if the true model is contained in the search list, the criterion will (in the limit for $n \to \infty$) select this correct model. For a related construction for factor models, see Bai and Ng (2002). This motivates us to take $C(n) = \log(n)$, and we define our second criterion

$$\text{SVMICb}(S) = \sum_{i=1}^{n} \xi_i + \log(n)|S|. \tag{9}$$

It is immediate that the computational cost of both SVMICs is much lower than of the cross-validated error rate (10 more SVMs to train for 10-fold cross-validation) and of the

kernel regularisation information criterion KRIC (which needs computations of the order $\mathcal{O}(n^3)$ due to the matrix inversion). The best case is when the $\xi_{i,S}$ are directly available. Computing the SVMICs is only an $\mathcal{O}(n)$ computation in that case, and usually even less when employing the property that

$$\xi_{i,S} \neq 0 \Leftrightarrow \alpha_{i,S} = 1.$$

When only α_S and Q_S are available, $\xi_{i,S}$ is computed using the relation

$$\xi_{i,S} = \Big[1 - y_i \sum_{\substack{j=1 \\ \alpha_{j,S}>0}}^{n} \alpha_{j,S}[Q_S]_{ij}\Big]_+.$$

This means that in the worst case, the computation time of the $SVMIC$s is $\mathcal{O}(n^2)$, which is still faster than using either CV error rate or KRIC.

4. Simulation Results

We perform $M = 100$ simulation runs with the following settings. We generate $n \in \{25, 50, 100, 200\}$ independent observations x_i, $1 \leq i \leq n$ of dimension $p \in \{25, 50, 100, 200\}$, with distribution $\mathcal{N}(0, \sigma^2 I_p)$ where $\sigma^2 = 1$. For each observation we generate a class label $y_i \in \{-1, +1\}$, with $P(y_i = 1) = 1/2$. Finally, we let $\mu = (1/2, -1/2, -1/2, 1/2, 0, \ldots, 0)$ of dimension p, and set $x_i \leftarrow x_i + y_i\mu$ to separate the two classes to some extent. This implies that the optimal separating hyperplane is $x'\mu = 0$, such that $\hat{y} = +1$ if $x'\mu > 0$, resulting in a generalization error rate of $\Phi(-\|\mu\|_2/\sigma)$, with Φ the cumulative distribution function of a standard normal. In our example, with $\sigma = 1$ and $\|\mu\|_2 = 1$, we find an optimal generalization error rate of 0.159.

During each simulation run, we standardize the variables to improve the numerical performance of the SVM algorithm. The variables are ranked using either the Fisher score or based on the variable influence on w, as described in Section 2.3. For each of the nested models obtained in the variable ranking step, we compute (i) SVMICa and (ii) SVMICb as in (8) and (9). We compare their performance to (iii) ten-fold CV, (iv) Vapnik's GRM as in (4), (v) KRIC for the logistic Bayesian model for SVMs as in (5), and (vi) KRIC for the Sollich model for SVMs as in (6). An important remark is that for ten-fold CV, we employ the CV2 method, which includes the feature selection procedure in each cross-validation step, as suggested by Zhang et al. (2006). Computing the CV error rate in the usual way can lead to a (severely) biased estimate of the generalization error, and using CV2 reduces this bias.

The experiment is repeated with two different kernels (i) a linear kernel $K(x_1, x_2) = x_1'x_2$ leading to a linear decision rule (ii) a quadratic kernel $K(x_1, x_2) = (\gamma x_1'x_2 + 1)^2$, with $\gamma = 1/p$, the inverse of the number of variables, leading to a quadratic decision rule. The tuning parameter C in each SVM that we train is chosen to be $C = 1$, as we standardize the explicative variables a priori. This is also the standard setting for C for the `svm` procedure in the `R` software package. We experimented with other values of C in the range from 0.1 up to 10, and found only minor differences in the simulation outcomes. We test the accuracy of the classifiers computed from the selected input variables by estimating their generalization

Linear kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	32.2	29.4	32.6	31.6	33.5	31.8	36.2	34.5	31.3	29.0	31.5	29.9
	50	34.6	31.6	35.3	32.6	35.3	33.5	37.4	35.4	34.4	33.2	34.4	33.2
	100	37.4	33.9	37.3	35.0	37.8	34.4	38.6	35.7	37.0	34.9	37.1	34.9
50	25	24.4	21.6	24.6	23.2	27.1	25.5	31.1	29.6	25.7	24.9	26.0	25.9
	50	28.5	23.3	27.7	24.8	29.5	26.3	31.4	30.5	29.8	28.7	30.2	29.7
	100	30.9	24.6	29.1	25.0	31.0	28.0	32.1	30.9	31.0	30.1	31.3	30.8
100	25	19.9	18.5	19.6	18.9	24.6	23.8	30.1	30.1	21.8	20.6	22.3	21.7
	50	22.9	19.2	20.2	19.0	25.8	25.4	29.9	29.6	26.9	26.8	27.3	27.8
200	25	17.8	17.0	16.9	16.8	22.7	21.5	28.9	29.3	18.7	18.0	19.2	18.9
Quadratic kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	31.3	30.7	34.2	33.8	33.8	32.9	37.7	36.6	29.5	28.4	30.2	30.1
	50	35.8	35.3	39.3	38.5	39.6	38.5	43.6	42.6	33.3	33.0	33.9	34.1
	100	43.3	43.3	48.3	48.4	42.8	42.7	49.2	48.7	37.1	37.1	37.7	38.2
50	25	22.7	21.3	25.0	24.3	26.7	25.9	31.8	31.7	23.6	22.5	24.8	25.1
	50	24.4	23.0	26.8	26.8	29.8	28.1	33.9	33.5	27.6	27.1	29.1	29.3
	100	26.4	25.6	30.8	30.2	34.1	33.8	40.3	40.1	31.1	30.9	32.5	32.8
100	25	19.4	18.5	19.9	19.1	23.8	19.2	30.6	30.2	20.0	20.0	21.7	22.0
	50	19.7	18.5	19.8	19.5	24.2	22.0	30.5	30.7	22.6	22.6	24.7	25.1
200	25	20.1	20.3	17.1	16.8	22.4	21.4	29.4	29.6	18.3	18.1	20.3	20.6

Table 1: Simulated average generalization error rate (%) for the six methods using two different kernels. For each method, the number on the left resulted from ranking by variable influence on $\|w\|^2$, and the number on the right in each column is from ranking by the Fisher scores S_j.

(out-of-sample) error rate from a test sample of 10000 new observations. These observations are generated in the same way as the training sample.

Table 1 reports the generalization error rates, obtained by averaging over the 100 simulation runs. An overall observation is that the error-rate based selection criteria (CV and GRM) have the worst performance. The performances of the KRICs and the new SVMICs are comparable. More precisely, we observe that the KRICs are better as a variable selection method for small sample sizes ($n = 25$), while the SVMICs give better results for larger sample sizes. This is especially apparent when the quadratic kernel is used. For a small number of observations compared to the number of variables, we also note that SVMICa slightly outperforms SVMICb in terms of generalization error rate, and that the opposite is true with many observations and fewer variables. The differences in generalization error rates become smaller as the number of variables grows. This is particulary true for CV, whose relative performance becomes better at large sample sizes. But SVMICa and SVMICb are still somewhat ahead, and have the advantage that they are much easier (and less time-intensive) to compute than the other criteria, included the KRICs having a computation time of order $\mathcal{O}(n^3)$. Note that, as n grows, the generalization error rates of the models obtained by our two suggested criteria are converging towards the theoretically obtained minimal generalization error rate of 15.9%. Investigating which variable ranking criterion is better, results in case of linear kernels to a strong preference for ranking with the Fisher score. For the quadratic kernel, it is slightly better to rank the variables based on variable influence on $\|w\|^2$.

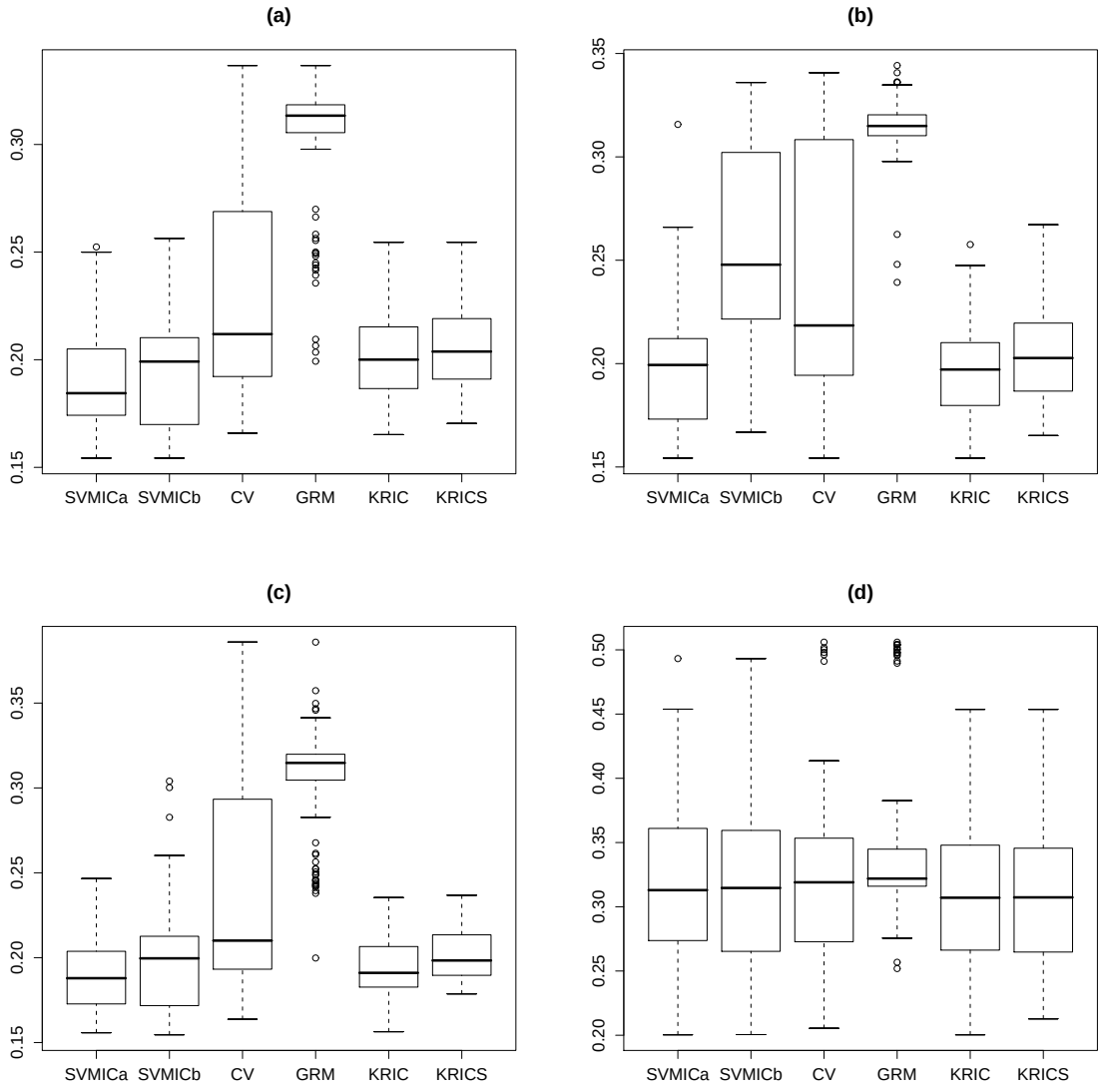

Figure 2: Generalization error rates for 100 simulation experiments, for $n = 100$, $p = 25$ (a) linear kernel, ranking with $\|w\|^2$, (b) linear kernel, ranking with Fisher score, (c) quadratic kernel, ranking with $\|w\|^2$, and for (d) $n = 25$, 100 variables, linear kernel and ranking with $\|w\|^2$.

Figure 2 presents the values of the 100 simulated generalization errors as boxplots, giving insight in the variability of the variable selection methods. For most of the cases it turns out that cross-validation is highly variable, while GRM has a small variability. This good property of GRM is, however, accompanied by a much higher average generalization error rate. Comparing the different information criteria shows that SVMICa is quite comparable to the KRICs. The SVMICb has a larger variability. In the setting with small sample size

($n = 25$) and relatively large number of variables (100), all methods, except for GRM, are comparable with respect to variability, but GRM has again the largest median error rate. Our main conclusion from this analysis is that SVMICa has a similar variability than the KRIC criteria, but SVMICb has a larger variability. Recall that the average error rates, as reported in Table 1, were of similar magnitude for all the four information criteria. Hence, when needing to choosing between the two newly proposed information criteria, we have a preference for SVMICa.

Given the variability of the generalization errors over the 100 simulation runs, see the boxplots in Figure 2, it is important to test whether the averages reported in Table 1 are also significantly different from each other. We performed standard t-tests, and most difference are indeed significant. For example, for the settings presented in Figure 1, we obtained that, at the 1% level, (a) all differences are significant, except between SVMICb and the 2 KRiCs (b) all differences are significant, except between SVMICa and the 2 KRICs (c) all differences are significant, except between SVMICb and the 2 KRICs (d) the differences with the GRM method are significant, the others not.

Furthermore, we investigate which models are actually chosen by the different criteria. This information is reported in Table 2. For each setting, it shows how many times the correct subset of input variables, containing only the first four input variables, was chosen (C, correct). This table also shows how many times a too-sparse group of variables was selected (U, underfitting), and how many times a too-rich group of variables was chosen (O, overfitting). So an overfit means that all correct variables are selected, but in addition some superfluous ones, while an underfit selects a subset of the important variables, but no irrelevant variables are included. The good performance of SVMICa and SVMICb might be due to the fact that these criteria seem to have the tendency to select a set of variables which includes all significant ones as the number of observations grows. The simulation results indicate that SVMICa behaves like AIC with its tendency to overfit. The SVMICb seems to share the property of BIC that it selects the correct model more often, if at least this true model is one of the possibilities to select from. The cross-validated error rate, and the general risk minimisation in particular, seem to have the tendency to ignore variables which nevertheless are important. As a consequence, the models that these criteria select are of poor predictive quality. The two KRICs of Kobayashi and Komaki (2006) share the overselection property exhibited by SVMICa, but the KRICs select excessive variables even more frequently than SVMICa. This can explain why these criteria perform somewhat worse when the number of observations is large, and why they outperform the proposed SVMICs when the number of observations is small, since the latter tend to underfit the model in the case of few observations.

This concludes the results for the case of two populations coming from an identical distribution, differing only in mean. Another case that we examined is where the variances of the two populations differ from each other. We performed a simulation study, in a similar way as the previous one, where the samples have been drawn from $\mathcal{N}(\mu, I_p)$ for class $+1$, and from $\mathcal{N}(-2\mu, 4I_p)$ for class -1.

The results of this simulation are summarized in Tables 3 and Table 4. We observe similar results as in the case where both populations had equal variance. Selection based on CV error rate and on GRM still perform rather poor. As before, the performances

	Kernel:	Linear				Quadratic			
	Models selected:	C	U	O	R	C	U	O	R
$n = 25$; $p = 25$	SVMICa	1	22	1	76	3	36	0	61
	SVMICb	0	42	0	58	0	64	0	36
	CV	0	38	4	58	1	40	5	54
	GRM	0	77	0	23	0	75	0	25
	KRIC	1	1	7	91	0	1	25	74
	KRICS	0	0	9	91	0	0	49	51
$n = 200$; $p = 25$	SVMICa	22	0	76	2	2	0	98	0
	SVMICb	77	9	10	4	67	14	6	13
	CV	7	48	43	2	4	43	49	4
	GRM	1	98	1	0	1	99	0	0
	KRIC	6	0	93	1	8	0	84	8
	KRICS	1	0	99	0	0	0	100	0
$n = 25$; $p = 100$	SVMICa	0	8	0	92	0	35	0	65
	SVMICb	0	20	0	80	0	63	0	37
	CV	0	23	6	71	0	33	10	57
	GRM	0	56	0	44	0	64	0	36
	KRIC	0	1	0	99	0	0	41	59
	KRICS	0	0	1	99	0	0	56	44

Table 2: Simulated frequencies of selected models, with variable ranking done by influence on $\|w\|^2$. Here 'C' denotes correct selection, 'U' is underfitting, 'O' is overfitting, and 'R' for all other situations.

Linear kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	28.9	28.0	30.1	29.2	30.4	28.4	32.7	31.6	29.0	27.5	28.8	27.7
	50	33.3	30.2	34.2	31.3	35.1	31.4	35.3	33.1	32.7	30.7	32.5	30.5
	100	35.6	31.5	35.7	32.3	36.0	32.6	36.9	33.7	34.8	32.6	34.8	33.0
	200	36.5	33.2	36.4	34.4	36.4	34.2	36.6	35.6	36.4	33.5	36.1	33.7
50	25	23.3	20.5	23.9	21.9	26.1	24.9	28.9	28.6	24.2	23.6	24.6	24.3
	50	27.1	21.7	25.7	22.7	27.7	25.2	29.1	28.4	27.7	26.8	27.6	27.1
	100	28.3	23.1	27.4	23.7	28.7	25.2	29.9	28.7	28.4	26.7	28.4	27.5
100	25	19.0	17.4	18.1	17.4	22.7	21.5	27.6	27.6	20.5	20.0	21.0	20.9
	50	21.8	17.8	19.3	18.0	23.5	22.7	26.9	27.0	24.8	25.0	25.0	25.5
200	25	17.0	16.1	15.9	15.6	21.4	20.7	27.0	27.0	17.9	17.0	18.3	17.8
Quadratic kernel													
n	p	SVMICa		SVMICb		CV		GRM		KRIC		KRICS	
25	25	29.2	28.9	31.8	31.8	31.8	28.7	35.4	34.7	25.7	24.9	25.8	26.2
	50	35.1	35.8	39.6	40.0	38.1	37.6	42.8	42.4	30.5	30.8	31.3	32.3
	100	42.1	41.7	48.2	48.1	42.2	42.3	49.4	48.7	35.0	36.0	36.2	38.1
	200	50.1	50.1	50.1	50.1	44.7	44.4	50.1	50.1	38.9	40.0	40.4	41.8
50	25	20.5	19.3	23.5	22.2	25.9	24.5	30.6	30.2	19.0	19.1	19.5	19.9
	50	23.1	22.2	26.1	26.2	28.3	27.6	33.2	32.7	23.8	23.9	25.1	26.1
	100	26.5	25.8	30.4	30.4	34.5	33.7	40.5	40.4	28.2	28.8	30.1	32.3
100	25	14.6	15.2	18.5	16.4	20.8	19.9	27.8	27.1	14.2	14.5	14.5	14.9
	50	17.9	17.0	18.4	17.8	22.0	21.5	27.7	28.3	18.1	18.5	19.5	20.3
200	25	9.9	9.8	12.9	13.2	19.6	17.6	29.3	26.8	10.1	10.3	9.7	9.8

Table 3: As Table 1, but now for two populations with different variances

of the KRICs and SVMICs are similar. More precisely, the SVMICs have an improved performance with respect to the KRICs when the sample size is large ($n \geq 50$) and the linear kernel is used, and the KRICs work slightly better for small sample sizes ($n = 25$). For the quadratic kernel, we notice a good performance of the KRICs, which is only matched

	Kernel:	Linear				Quadratic			
	Models selected:	C	U	O	R	C	U	O	R
$n = 25$; $p = 25$	SVMICa	0	22	1	77	1	36	0	63
	SVMICb	0	47	0	53	1	57	0	42
	CV	1	40	1	58	1	39	8	52
	GRM	0	76	0	24	0	70	0	30
	KRIC	0	0	6	94	0	0	25	75
	KRICS	0	0	8	92	0	0	50	50
$n = 200$; $p = 25$	SVMICa	11	0	85	4	0	20	0	80
	SVMICb	69	10	16	5	0	45	0	55
	CV	6	56	37	1	0	33	4	63
	GRM	0	100	0	0	0	56	0	44
	KRIC	5	0	93	2	0	0	40	60
	KRICS	0	0	99	1	0	0	53	47
$n = 25$; $p = 200$	SVMICa	0	1	0	99	0	52	0	48
	SVMICb	0	8	0	92	0	54	0	46
	CV	0	22	2	76	0	22	5	73
	GRM	0	46	0	54	0	54	0	46
	KRIC	0	1	0	99	0	0	46	54
	KRICS	0	0	0	100	0	0	56	44

Table 4: As Table 2, but now for two populations with different variances

by SVMICa for larger sample sizes. From Table 4 we can again make the same observations as before when the linear kernel is used. For the quadratic kernel the SVMICs have more difficulty selecting all the relevant variables than the KRICs, which explains why the latter criteria have an improved performance here.

We also conducted a simulation experiment where the input variables were strongly correlated. First, the observations were generated as in the first simulation experiment. Then, we applied the transformation

$$x_{ij} = \rho x_{ik_j} + \epsilon_{ij} \text{ with } \epsilon_{ij} \sim \mathcal{N}(0, \rho^2) \text{ i.i.d.}$$

where $i = 1, \ldots, n$, k_j is chosen arbitrarily between 1 and 4, and $4 < j \leq p/2$, such that about half of the unimportant input variables are correlated with the four important ones. The parameter $|\rho| < 1$ controls the degree of correlation. We have chosen $\rho = 0.8$ and found similar results (not reported) as for the case where the variances of both class-population differ.

5. Tests on Real Data Sets

We compare the performance of the new methods with that of the other discussed criteria on several real-world datasets. We use some of the benchmark datasets used in Rakotomamonjy (2003), and in Rätsch et al. (2001). The datasets used are the Pima Indians Diabetes database (768 observations, 8 variables), the Statlog Cleveland Heart Disease database (303 observations, 14 variables), and Leo Breiman's ringnorm and twonorm datasets (both 7400 observations, 20 variables). These datasets are available from the UCI Machine Learning Repository (the first two), and the Delve Repository (last two). We perform 100 random splits of the data in a training sample and a test sample, where the size of the training sample is chosen as $\sqrt{2n}$, with n the total number of observations in the dataset. We chose the size of the training set such that there is a sufficient amount of observations in the test

	Ranking:	Variable influence on $\|w\|$			Fisher scores		
Data	Kernel:	Linear	Quadratic	Radial	Linear	Quadratic	Radial
Diabetes	SVMICa	28.6	28.5	29.2	28.0	28.2	28.4
	SVMICb	29.0	28.9	29.2	28.6	28.5	28.9
	CV	28.6	29.1	29.1	28.8	28.5	29.3
	GRM	29.6	29.7	29.6	29.1	29.2	29.3
	KRIC	28.5	28.2	29.4	27.5	28.1	29.6
	KRICS	28.6	28.5	29.7	28.3	28.6	29.7
Heart	SVMICa	27.0	27.4	27.7	27.6	28.0	28.3
	SVMICb	27.6	28.9	28.9	28.2	29.3	29.5
	CV	27.6	28.6	27.2	26.8	28.0	28.8
	GRM	29.3	30.3	29.4	28.8	30.4	30.6
	KRIC	25.4	23.4	23.8	24.5	23.2	23.8
	KRICS	25.3	23.5	25.2	25.2	23.7	25.0
Ringnorm	SVMICa	31.1	16.4	8.4	30.8	15.6	6.5
	SVMICb	34.9	20.2	13.5	35.2	22.4	13.4
	CV	33.9	32.1	26.6	32.8	25.6	21.2
	GRM	39.2	41.3	38.6	39.3	38.4	37.3
	KRIC	30.1	16.3	6.0	29.6	15.9	4.4
	KRICS	29.9	16.0	3.1	29.2	15.4	2.5
Twonorm	SVMICa	9.9	9.3	11.4	10.1	8.9	9.4
	SVMICb	13.5	14.1	15.9	15.0	15.2	16.0
	CV	20.5	21.0	19.8	21.0	21.1	20.8
	GRM	31.4	31.7	31.6	30.8	31.2	31.3
	KRIC	8.0	7.5	11.0	6.8	6.8	9.2
	KRICS	7.5	6.0	4.0	6.6	5.5	4.8

Table 5: Generalization error rates (%) for variable selection applied to four data sets. Two variable ranking schemes and three types of kernel are used for each of the criteria.

sample to estimate the generalization (out-of-sample) error rate. The training sample size is relatively small, such that the computation time for the KRIC remains within bounds. For each of these partitions we perform variable selection on the training sample exactly as in the simulation study. We first rank the variables to retain p stacked subsets of input variables, and then use the information criteria to select the variables that best explain the training data. Then, we predict the class labels for the test sample, and use these predictions to estimate the generalization error rate. We use variable ranking based on variable influence on $\|w\|^2$ as well as on Fisher score, and we use a linear, quadratic and radial kernel.

The estimated generalization error rates are presented in Table 5 for each dataset and estimation setting. We observe that the KRICs are the preferred choice of variable selection criterion in terms of generalization error rate for the 'twonorm' and 'heart' datasets. For the 'ringnorm' and 'diabetes' datasets the difference in performance between the KRICs and our newly proposed SVMICs is less pronounced. The predictive performance of the models selected by SVMICa are for most settings comparable to that of the KRIC, while being much faster to compute. These results are consistent across all settings. The CV error rate and especially the GRM have a poor performance, which is in line of the results obtained in the simulation.

From these results, and the results obtained in Section 4, we suggest to use either the SVMICa or the SVMICb if a preliminary analysis of the data or a priori knowledge

indicates that the true decision function is almost linear. When it differs strongly from a linear function, the researcher has a choice between the ease of computation of the support vector machine information criteria, or the somewhat improved predictive performance, though with higher computational cost, of the kernel regularization information criterion.

Finally, we applied the newly proposed information criteria for variable selection to two large data sets, the "Madelon" ($n = 2000, p = 500$) and "Arcene" data ($n = 100, p = 10000$). These data sets were part of the NIPS 2003 feature selection, and are described in detail in Guyon et al (2006). Given the high dimensionality of these data, the variables were ranked according to the Fisher score. We used a linear kernel and computed balanced error rates (BER), that is the average of the error rate of the positive class and the error rate of the negative class. When using SVMICa we obtain a BER of 43.0% for the Madelon data, and 31.1% for the Arcene data. For SVMICb we get 37.3% and 31.1%, respectively. In Guyon et al (2006, 2007) the BER of other feature selection methods is presented, and it turns out that several other methods yield much better performance on these data. A possible explication is that we used a standard SVM, without any optimal tuning of the regularization parameters.

6. Conclusions

In this paper we considered the problem of variable selection in support vector machines. We proposed two new information criteria, SVMICa and SVMICb, which allow us to evaluate the suitability of the selected subset of variables for predictive purposes, without much additional computational costs. We provided an argumentation for these criteria, linking SVMICa to the KRIC of Kobayashi and Komaki (2006), and justifying SVMICb with the need for a consistent selection criterion. We demonstrated the effectiveness of these criteria in a simulation study, where we compared their predictive performance to the KRIC, cross-validation and general risk minimization. Especially for decision functions which are close to an affine function, we found that SVMICa and SVMICb performed the best of all tested criteria, and were also the easiest to compute. For more complicated decision functions, we found that SVMICa still performs well for selecting models with good generalization properties. We repeated the experiment on several real data examples, and the result confirmed the good properties of these newly proposed criteria. In particular we showed that cross-validation criteria are outperformed in generalization error by the new information criteria, where the latter are coming at almost no additional computational cost.

The aim of our paper was to propose an information criterion for a standard SVM. We do not claim that the procedure is outperforming other very advanced feature selection methods, which are not relying on a standard SVM. Obtaining information criteria for other machine learners is an interesting topic for future research. Another research question is how suitable the information criteria are for optimal tuning of the regularization and other parameters of the SVM, without necessarily selecting a subset of input variables. Finally, it would be interesting to continue on the theoretical verification of the good performance of our two proposed criteria, and for example try to obtain consistency results for the SVM information criteria.

Acknowledgments

We thank Professor Guyon and Professor Alamdari, the editors of the special issue of the Journal Machine Learning Research on model selection, and the reviewers for their constructive and insightful comments on the first submitted version of this paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, (eds. B. Petrov and F. Csáki), pages 267–281, Akadémiai Kiadó, Budapest, 2003.

Bai J. and Ng, S. Determining the number of factors in approximate factor models. *Econometrica*, 70: 191–221, 2002.

Bi J., Bennett K. P., Embrechts M., Breneman C. M. and Song, M. Dimensionality Reduction via Sparse support vector machines. *Journal of Machine Learning Research*, 3: 1229–1243, 2003.

Chen S.-W., Li Z.-R. and Li X.-Y. Prediction of antifungal activity by support vector machine approach. *Journal of molecular structure: THEOCHEM*, 731: 73–81.

Cristianini N. and Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, 2000.

Claeskens, G. and Hjort, N. L. *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, in press.

Fortuna J. and Capson D. Improved support vector classification using PCA and ICA feature space modification. *Pattern Recognition*, 37: 1117–1129, 2004.

Guyon I. and Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3: 1157–1182, 2003.

Guyon I., Gunn S., Nikravesh M., and Zadeh L. *Feature Extraction, Foundations and Applications.* Physica-Verlag, Springer, Berlin, 2006.

Guyon I., Li J., Mader T., Pletscher P. A., Schneider G., and Uhr M. Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters*, 28: 1438–1444, 2007.

Guyon I., Weston J., Barnhill S. and Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46: 389–422, 2002.

Hastie T., Tibshirani R. and Friedman J. *The elements of statistical learning: data mining, inference, and prediction.* Springer, New York, 2001.

Haughton, D. On the choice of a model to fit data from an exponential family, *The Annals of Statistics*, 16: 342–355, 1988.

Haughton, D. Size of the error in the choice of a model to fit data from an exponential family, *Sankhyā, Series A*, 51: 45–58, 1989.

Kearns M., Mansour Y., NG A. Y. and Ron D. An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning*, 27: 7–50, 1997.

Kobayashi K. and Komaki F. Information Criteria for support vector machines. *IEEE Transactions on Neural Networks*, 17: 571–577, 2006.

Lee Y., Kim Y., Lee S., and Koo J.-Y. Structured multicategory support vector machines with analysis of variance decomposition. *Biometrika*, 93: 555–571, 2006.

Lin Y. and Zhang H. H. Component Selection and Smoothing in Multivariate Nonparametric Regression. *Annals of Statistics*, 34: 2272–2297, 2006.

Neumann J., Schnörr C. and Steidl G. Combined SVM-Based Feature Selection and Classification. *Machine Learning*, 61: 129–150, 2005.

Peng H., Long F. and Ding C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 27: 1226–1238, 2005.

Rakotomamonjy A. Variable Selection Using SVM-based Criteria. *Journal of Machine Learning Research*, 3: 1367–1370, 2003.

Rätsch G., Onoda T. and Müller K.-R. Soft Margins for AdaBoost. *Machine Learning*, 42: 287–320, 2001.

Rissanen J. *Stochastic complexity in statistical inquiry*, *World Scientific* Series in Computer Science, volume 15. World Scientific, Singapore, 1989.

Schölkopf B. and Smola A.J. *Learning with Kernels*. MIT Press, Cambridge, 2002.

Schwarz G. Estimating the dimension of a model, *The Annals of Statistics*, 6: 461–464, 1978.

Shih F. Y. and Cheng S. Improved feature reduction in input and feature spaces. *Pattern Recognition*, 38: 651–659, 2005.

Sollich P. Bayesian Methods for support vector machines: evidence and predictive class probabilities. *Machine Learning*, 46: 21–52, 2002.

Woodroofe M. On model selection and the arc sine laws. *The Annals of Statistics*, 10: 1182–1194, 1982.

Vapnik V. N. *Estimation of dependences based on empirical data*. Springer, New York, 1982.

Wang L., Zhu J. and Zou H. The doubly regularized support vector machine. *Statistica Sinica*, 16: 589–615, 2006.

Zhang H. H. (2006). Variable Selection for SVM via Smoothing Spline ANOVA. *Statistica Sinica*, 16: 659–674, 2006.

Zhang X., Lu X., Shi Q., Xu X.-Q., Leung H.-C. E., Harris L. N., Iglehart J. D., Miron A., Liu, J. S., and Wong W. H. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, published 10 April 2006.

Zhu J., Rosset S., Hastie T. and Tibshirani R. 1-norm support vector machines. *Neural Information Processing Systems*, 16, 2004.